WWU MÜNSTER

Detection and visualization of complex structural variants from NGS data

Valerio Vitali
Institute for Evolution and Biodiversity
University of Münster, Germany

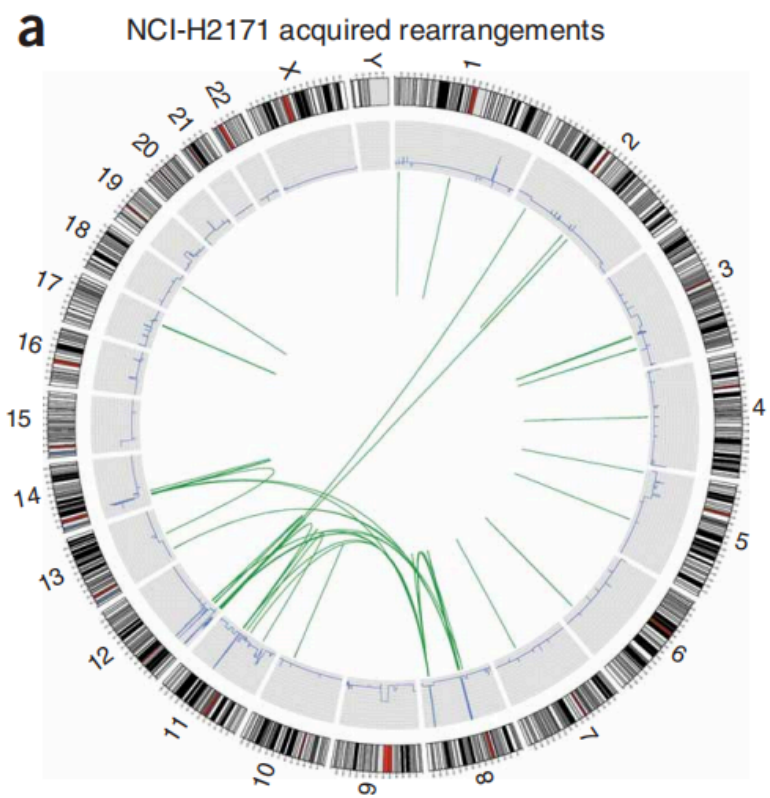
BIOINFORMATICS

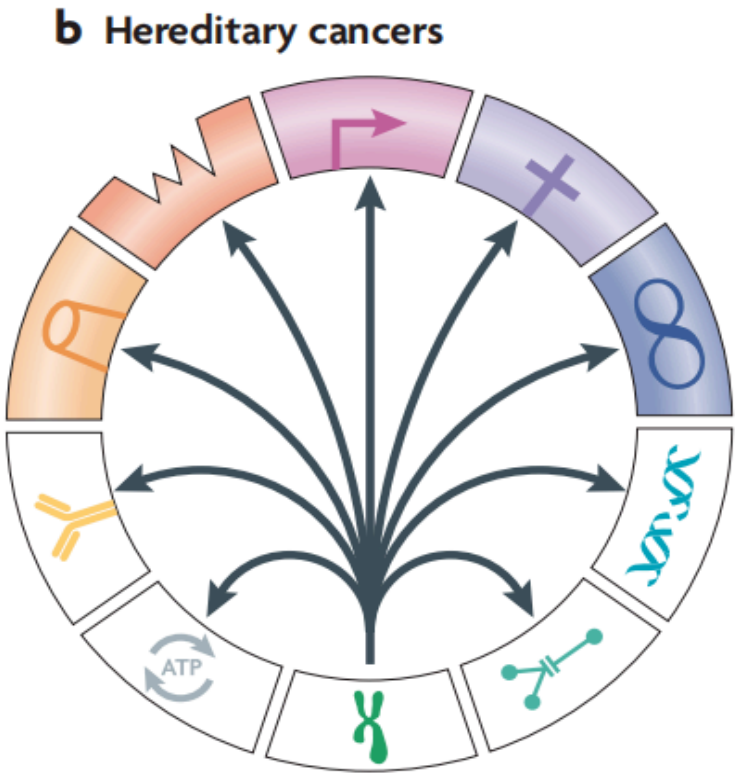
SESSION ON VARIANT CALLING

RTG 2220 EvoPAD

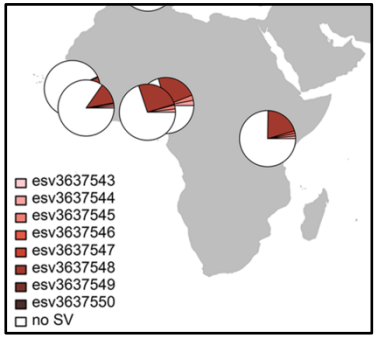
Summer School 2019

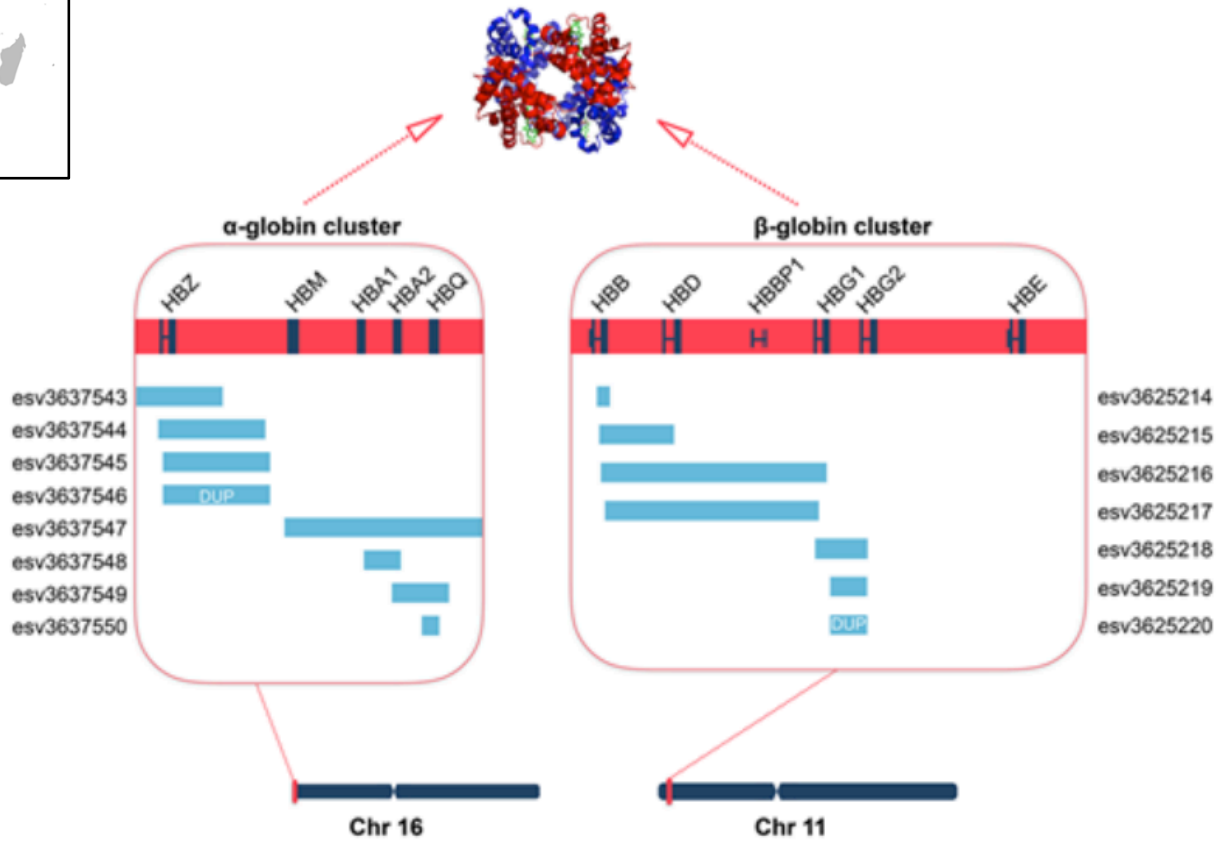
Structural Variants: initiation of cancer



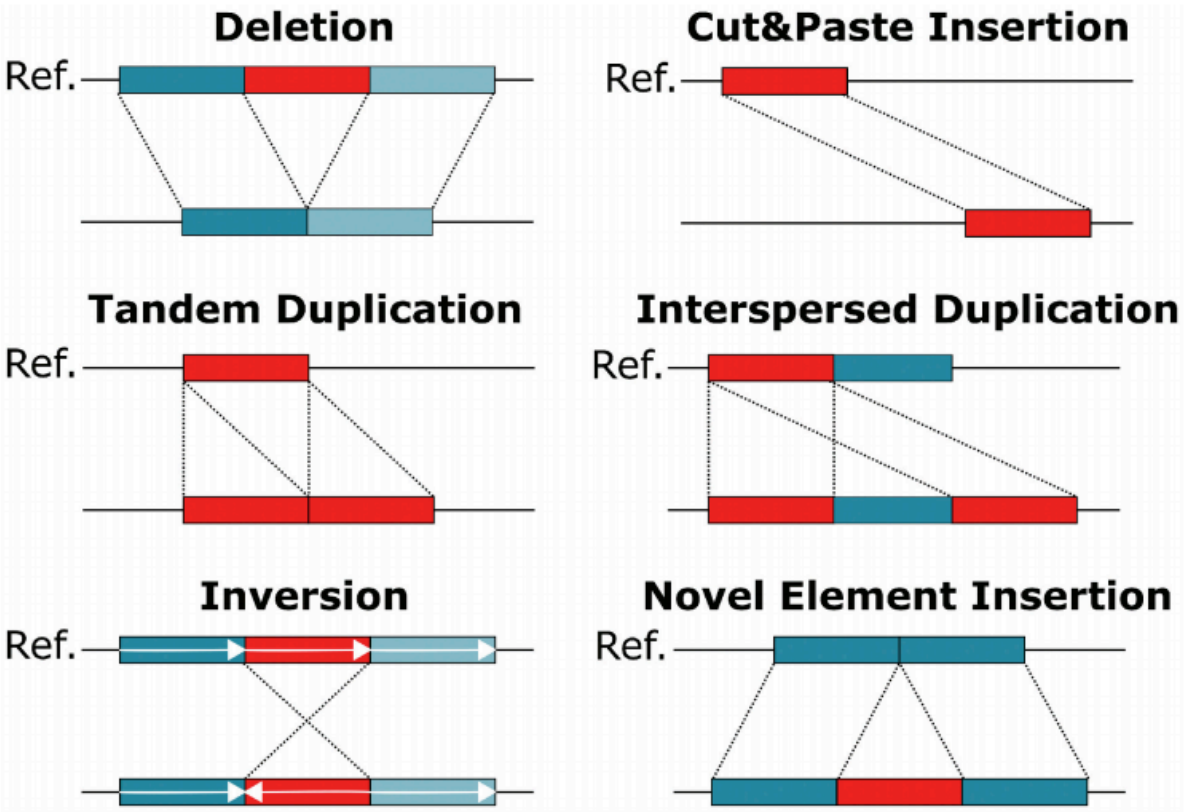


Structural Variants: balancing selection





Structural Variants: types

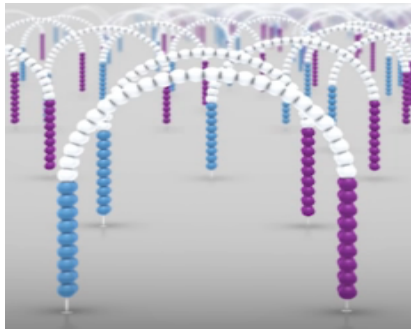


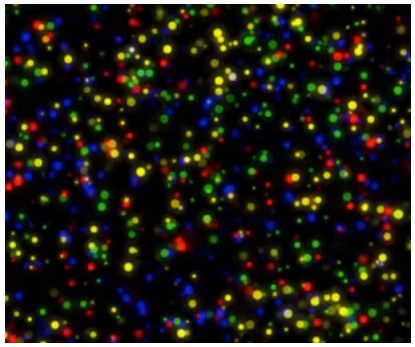
Illumina SBS vs PacBio CCS

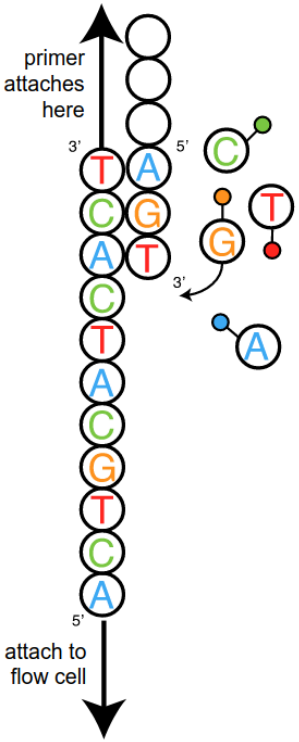
Illumina
Sequencing by Synthesis

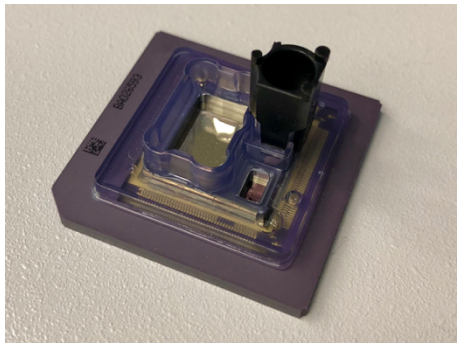
VS

PacBio
Circular Consensus Sequencing

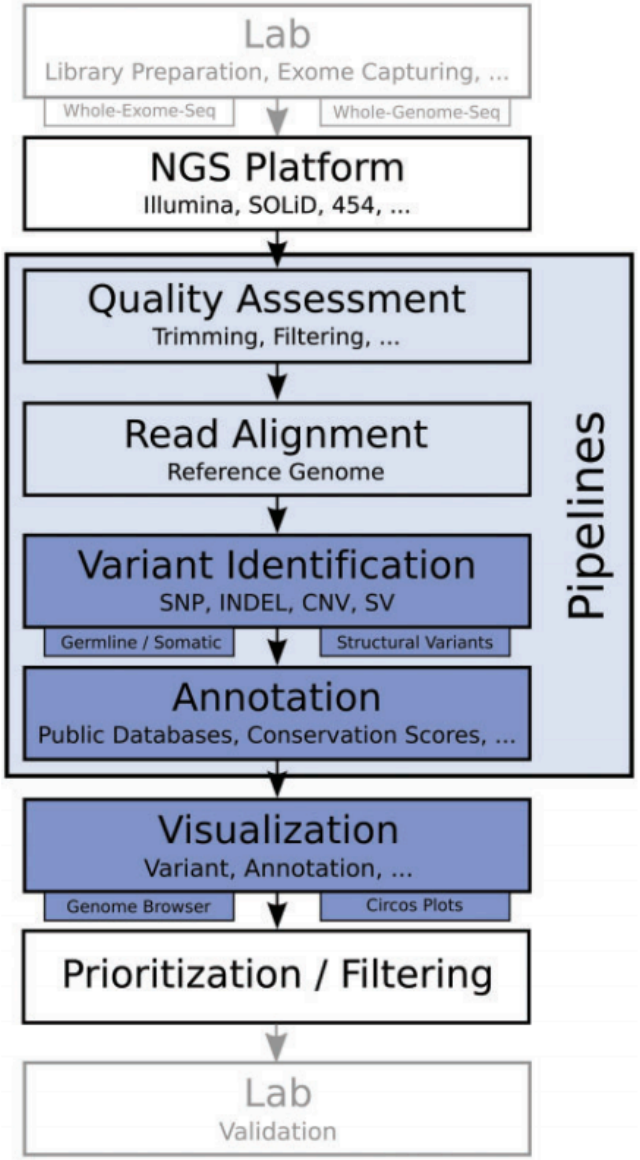








Standard NGS workflow



SV-CALLERS

Long Reads	Short Reads
• **Svim** • Sniffle • PBHoney	• **Pilon** • Delly • NovoBreak

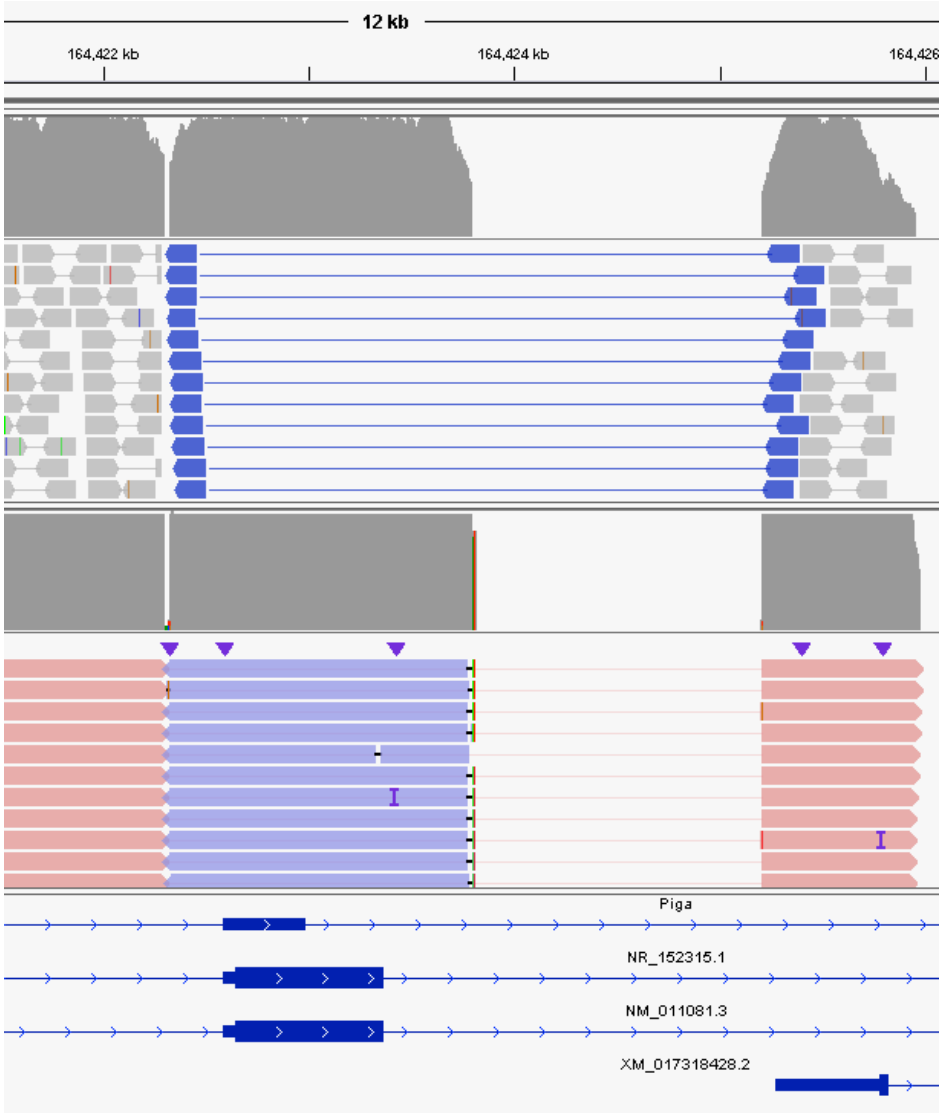
Short vs Long reads

+ Higher base quality
+ Lower costs

– **Indirect evidence for SVs**
– Mapping on repeated regions

+ Accurate mapping on repeated and low complexity regions
+ **Often span the whole SV**

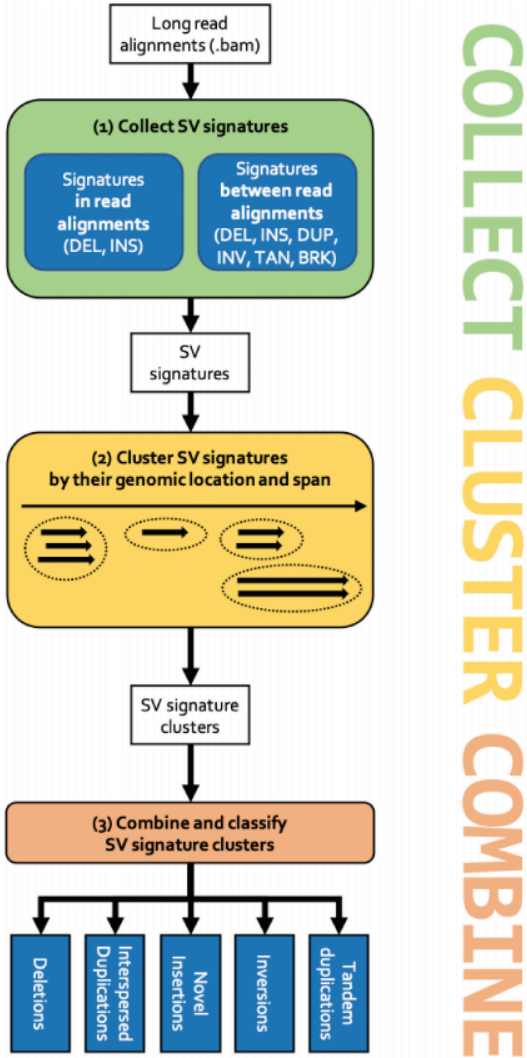
– Higher sequencing costs
– Elevated error rate (5-15%)



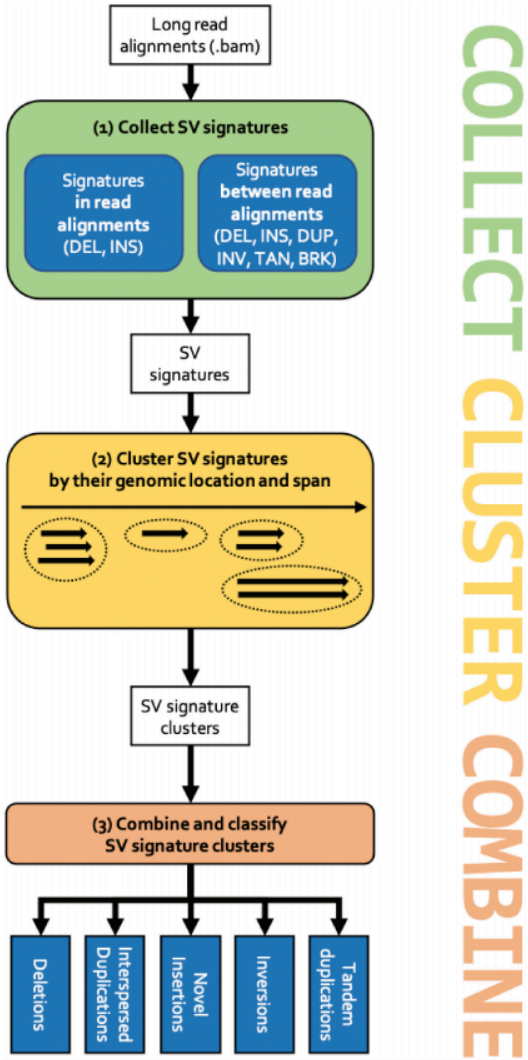
SV calling from PE-reads: Pilon

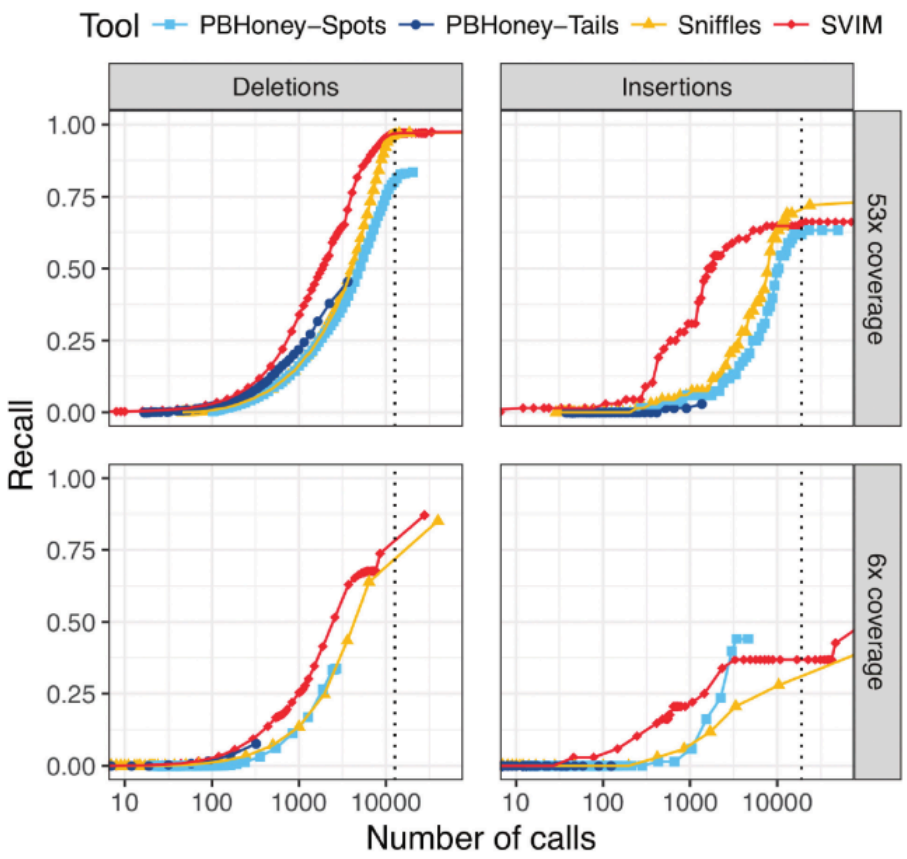
PROCESS	RESULT	
Pilon protocol	Assembly improvement (Fasta)	Variation detection (VCF)
Evaluate alignment pileups TAATGGGGGCGGTGCCATATCATGAGA TAATGGGGGCGGTGCCATATCATGAGA TAATGGGG•CGGTGCCATATCTAGAGA TAATGGGGGCGGTGCCATATCATGAGA	Identify and fix base errors	Identify SNPs and small indels
Scan read coverage and alignment discrepancies	Identify potential local misassemblies	Identify larger insertions and deletions
Reassemble across gaps and discrepant regions	Attempt to fill gaps and fix local misassemblies	Attempt to build out the full sequence of larger insertions

SV calling from long reads: svim



SV calling from long reads: svim





Variant Call Format (VCF)

Header

```
##fileformat=VCFv4.1
##fileDate=20190909
##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500"
##PILON="--genome GRCm38.p6_ChrX.fa --frags PIGA15ILL003.trmd.sorted.bam --output PIGA15ILL003 --outdir ./PIGA15ILL003.Pilon_output --changes --vcf --targets
NC_000086.7:164420377-164425982 --fix all,breaks --mindepth 3"
##reference=file:/home/waltermint/Desktop/PIPE1_test/GRCm38.p6_ChrX.fa
##contig=<ID=NC_000086.7,length=5606>
##FILTER=<ID=LowCov,Description="Low Coverage of good reads at location">
##FILTER=<ID=Amb,Description="Ambiguous evidence in haploid genome">
##FILTER=<ID=Del,Description="This base is in a deletion or change event from another record">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Valid read depth; some reads may have been filtered">
##INFO=<ID=TD,Number=1,Type=Integer,Description="Total read depth including bad pairs">
##INFO=<ID=PC,Number=1,Type=Integer,Description="Physical coverage of valid inserts across locus">
##INFO=<ID=BQ,Number=1,Type=Integer,Description="Mean base quality at locus">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Mean read mapping quality at locus">
##INFO=<ID=QD,Number=1,Type=Integer,Description="Variant confidence/quality by depth">
##INFO=<ID=BC,Number=4,Type=Integer,Description="Count of As, Cs, Gs, Ts at locus">
##INFO=<ID=QP,Number=4,Type=Integer,Description="Percentage of As, Cs, Gs, Ts weighted by Q & MQ at locus">
##INFO=<ID=IC,Number=1,Type=Integer,Description="Number of reads with insertion here">
##INFO=<ID=DC,Number=1,Type=Integer,Description="Number of reads with deletion here">
##INFO=<ID=XC,Number=1,Type=Integer,Description="Number of reads clipped here">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Fraction of evidence in support of alternate allele(s)">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVLEN,Number=.,Type=String,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise change from local reassembly (ALT contains Ns)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=String,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=String,Description="Approximate read depth; some reads may have been filtered">
##ALT=<ID=DUP,Description="Possible segmental duplication">
#CHROM  POS      ID        REF     ALT     QUAL     FILTER  INFO     FORMAT  SAMPLE
NC_000086.7      164420377        .       C       .        0        LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=1;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
     0/0
NC_000086.7      164420378        .       A       .        0        LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
     0/0
NC_000086.7      164420379        .       T       .        0        LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
     0/0
NC_000086.7      164420380        .       G       .        0        LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=3;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
```

Variant Call Format (VCF)

```
##fileformat=VCFv4.1
##fileDate=20190909
##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500"
##PILON="--genome GRCm38.p6_ChrX.fa --frags PIGA15ILL003.trmd.sorted.bam --output PIGA15ILL003 --outdir ./PIGA15ILL003.Pilon_output --changes --vcf --targets
NC_000086.7:164420377-164425982 --fix all,breaks --mindepth 3"
##reference=file:/home/waltermint/Desktop/PIPE1_test/GRCm38.p6_ChrX.fa
##contig=<ID=NC_000086.7,length=5606>
##FILTER=<ID=LowCov,Description="Low Coverage of good reads at location">
##FILTER=<ID=Amb,Description="Ambiguous evidence in haploid genome">
##FILTER=<ID=Del,Description="This base is in a deletion or change event from another record">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Valid read depth; some reads may have been filtered">
##INFO=<ID=TD,Number=1,Type=Integer,Description="Total read depth including bad pairs">
##INFO=<ID=PC,Number=1,Type=Integer,Description="Physical coverage of valid inserts across locus">
##INFO=<ID=BQ,Number=1,Type=Integer,Description="Mean base quality at locus">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Mean read mapping quality at locus">
##INFO=<ID=QD,Number=1,Type=Integer,Description="Variant confidence/quality by depth">
##INFO=<ID=BC,Number=4,Type=Integer,Description="Count of As, Cs, Gs, Ts at locus">
##INFO=<ID=QP,Number=4,Type=Integer,Description="Percentage of As, Cs, Gs, Ts weighted by Q & MQ at locus">
##INFO=<ID=IC,Number=1,Type=Integer,Description="Number of reads with insertion here">
##INFO=<ID=DC,Number=1,Type=Integer,Description="Number of reads with deletion here">
##INFO=<ID=XC,Number=1,Type=Integer,Description="Number of reads clipped here">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Fraction of evidence in support of alternate allele(s)">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVLEN,Number=.,Type=String,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise change from local reassembly (ALT contains Ns)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=String,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=String,Description="Approximate read depth; some reads may have been filtered">
##ALT=<ID=DUP,Description="Possible segmental duplication">
#CHROM  POS      ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
NC_000086.7     164420377       .       C       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=1;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420378       .       A       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420379       .       T       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420380       .       G       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=3;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
```

Columns

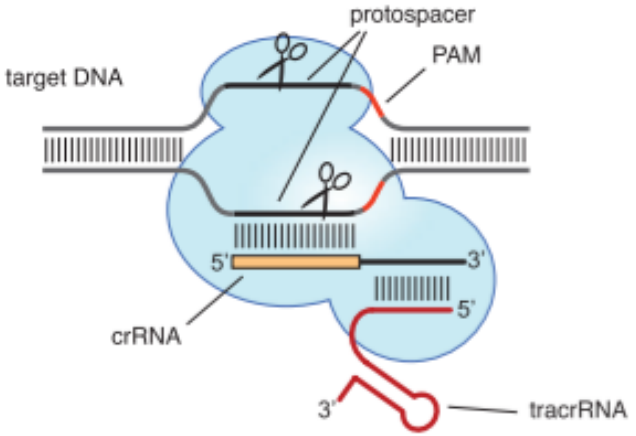
Variant Call Format (VCF)

```
##fileformat=VCFv4.1
##fileDate=20190909
##source="Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500"
##PILON="--genome GRCm38.p6_ChrX.fa --frags PIGA15ILL003.trmd.sorted.bam --output PIGA15ILL003 --outdir ./PIGA15ILL003.Pilon_output --changes --vcf --targets
NC_000086.7:164420377-164425982 --fix all,breaks --mindepth 3"
##reference=file:/home/waltermint/Desktop/PIPE1_test/GRCm38.p6_ChrX.fa
##contig=<ID=NC_000086.7,length=5606>
##FILTER=<ID=LowCov,Description="Low Coverage of good reads at location">
##FILTER=<ID=Amb,Description="Ambiguous evidence in haploid genome">
##FILTER=<ID=Del,Description="This base is in a deletion or change event from another record">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Valid read depth; some reads may have been filtered">
##INFO=<ID=TD,Number=1,Type=Integer,Description="Total read depth including bad pairs">
##INFO=<ID=PC,Number=1,Type=Integer,Description="Physical coverage of valid inserts across locus">
##INFO=<ID=BQ,Number=1,Type=Integer,Description="Mean base quality at locus">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Mean read mapping quality at locus">
##INFO=<ID=QD,Number=1,Type=Integer,Description="Variant confidence/quality by depth">
##INFO=<ID=BC,Number=4,Type=Integer,Description="Count of As, Cs, Gs, Ts at locus">
##INFO=<ID=QP,Number=4,Type=Integer,Description="Percentage of As, Cs, Gs, Ts weighted by Q & MQ at locus">
##INFO=<ID=IC,Number=1,Type=Integer,Description="Number of reads with insertion here">
##INFO=<ID=DC,Number=1,Type=Integer,Description="Number of reads with deletion here">
##INFO=<ID=XC,Number=1,Type=Integer,Description="Number of reads clipped here">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Fraction of evidence in support of alternate allele(s)">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVLEN,Number=.,Type=String,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise change from local reassembly (ALT contains Ns)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=.,Type=String,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=String,Description="Approximate read depth; some reads may have been filtered">
##ALT=<ID=DUP,Description="Possible segmental duplication">
#CHROM  POS      ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
NC_000086.7     164420377       .       C       .       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=1;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420378       .       A       .       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420379       .       T       .       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=2;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
    0/0
NC_000086.7     164420380       .       G       .       .       0       LowCov  DP=0;TD=0;BQ=0;MQ=0;QD=0;BC=0,0,0,0;QP=0,0,0,0;PC=3;IC=0;DC=0;XC=0;AC=0;AF=0.00 GT
```

INFO fields

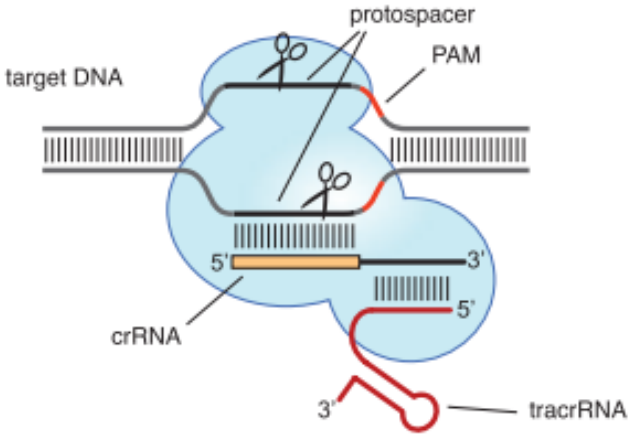
CRISPR-Cas9

Cas9 programmed by crRNA:tracrRNA duplex

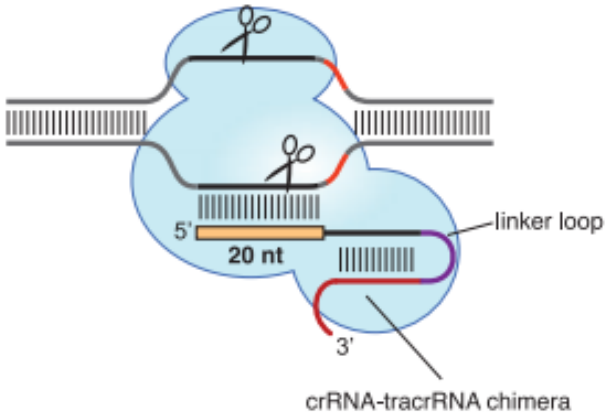


CRISPR-Cas9

Cas9 programmed by crRNA:tracrRNA duplex

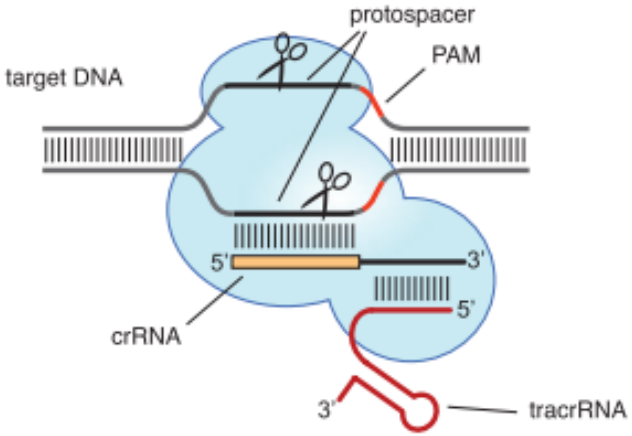


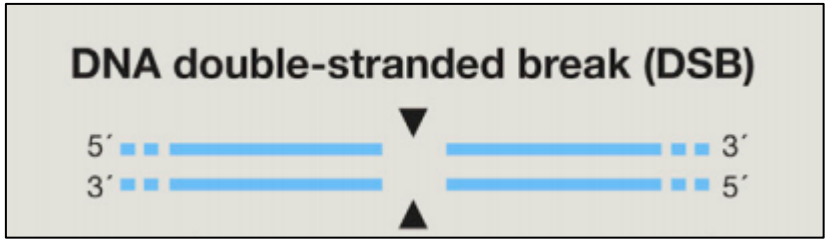
Cas9 programmed by single chimeric RNA



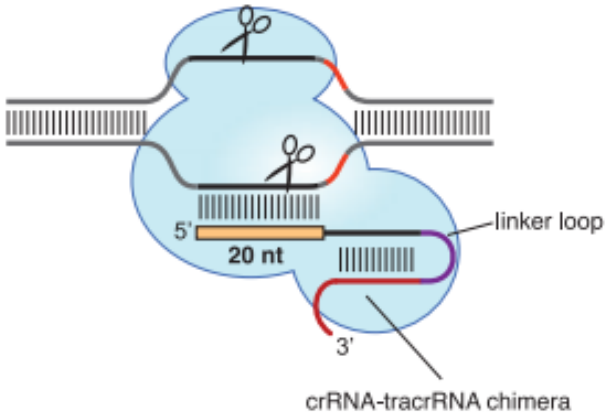
CRISPR-Cas9

Cas9 programmed by crRNA:tracrRNA duplex



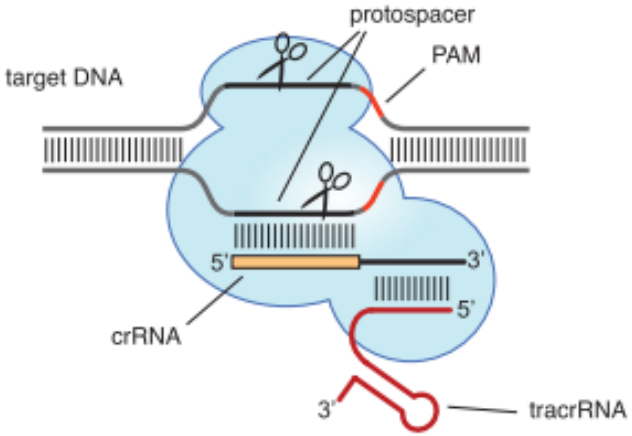


Cas9 programmed by single chimeric RNA

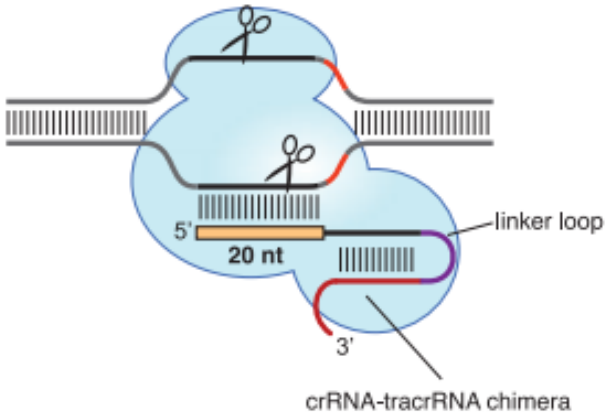


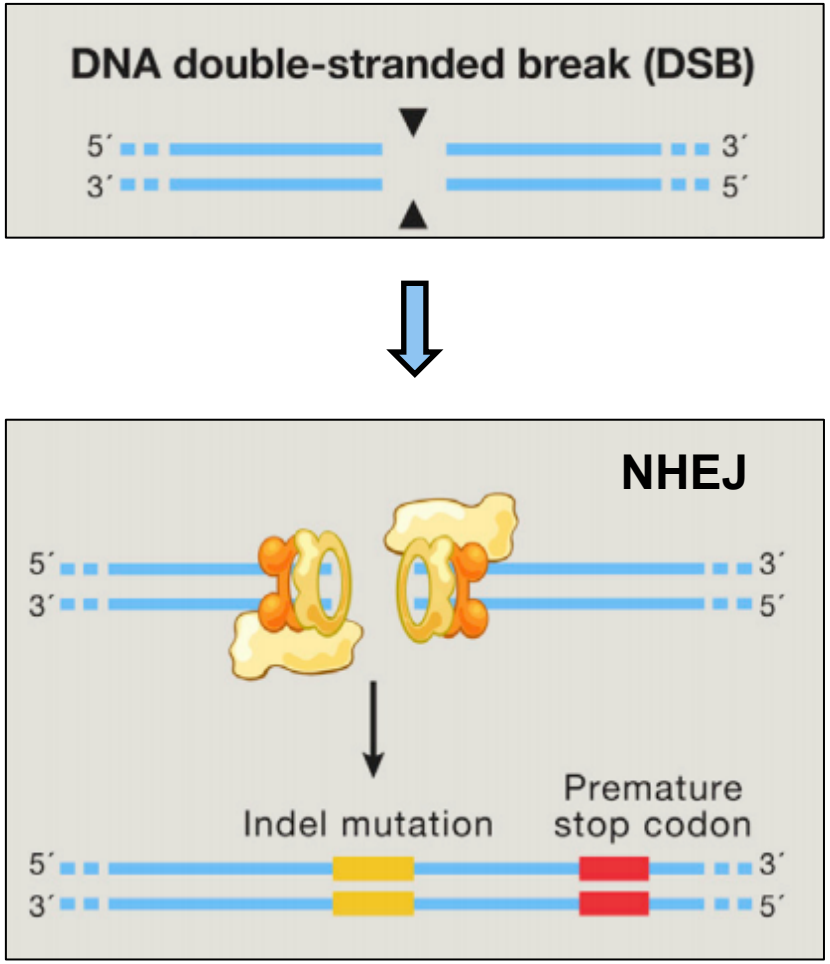
CRISPR-Cas9

Cas9 programmed by crRNA:tracrRNA duplex

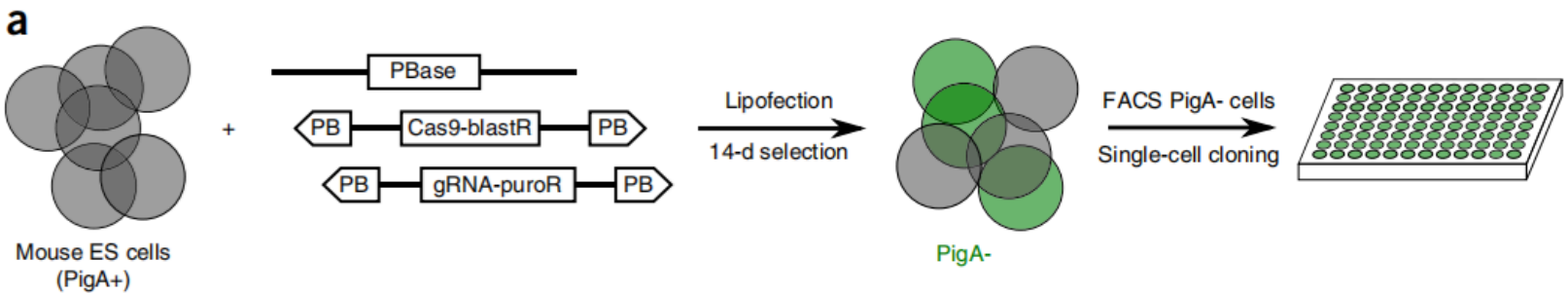


Cas9 programmed by single chimeric RNA

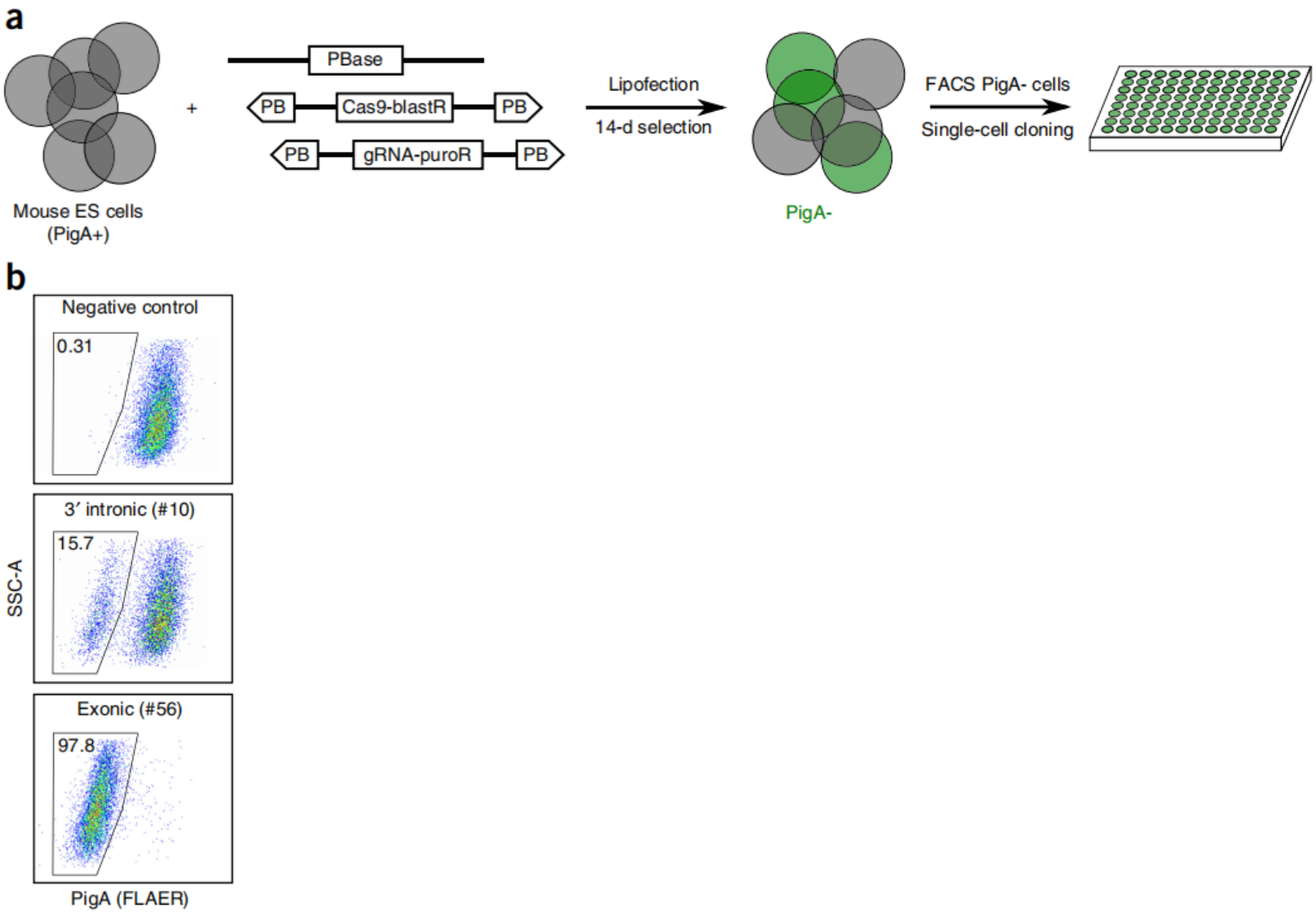




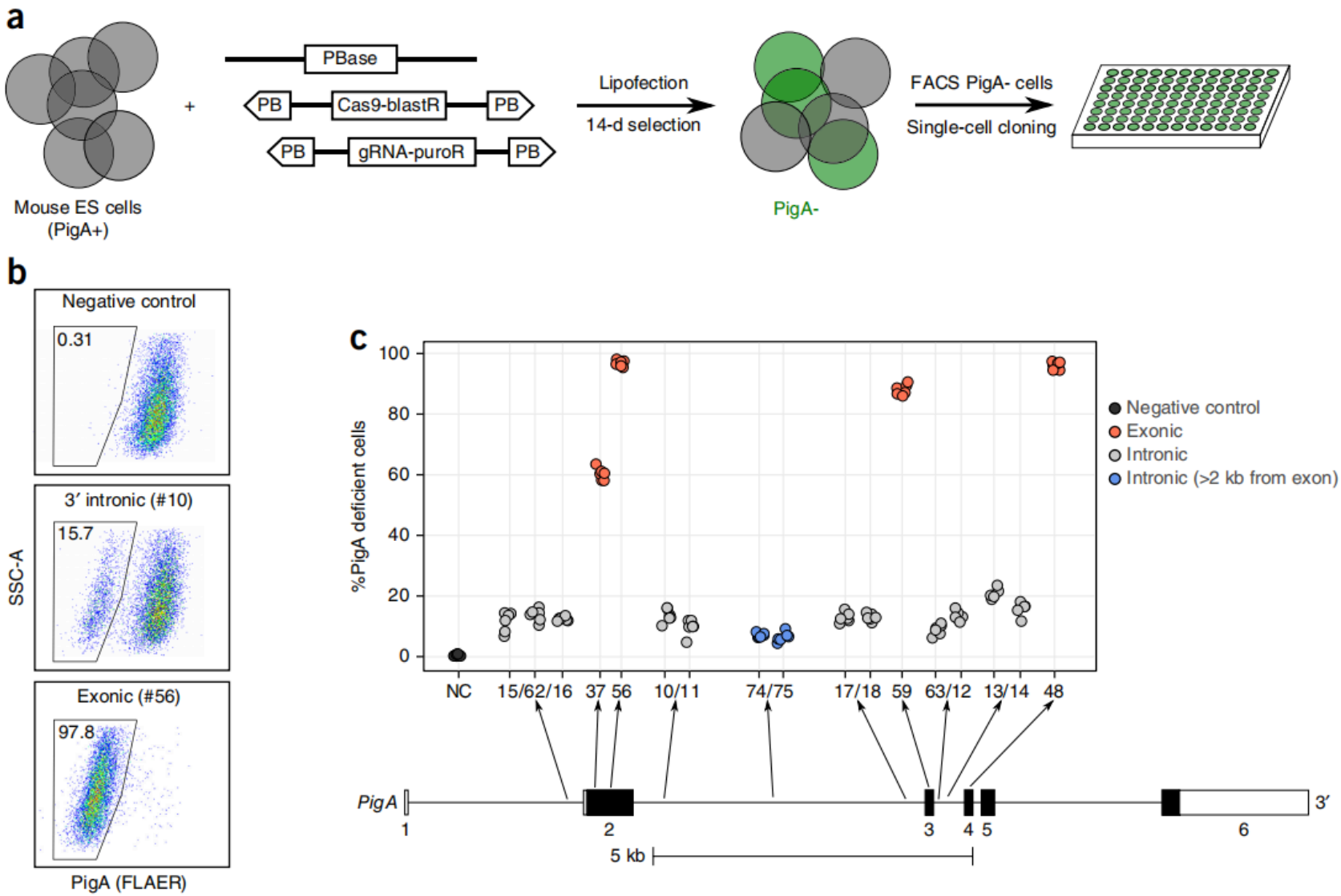
CRISPR-Cas9 leads to complex SVs



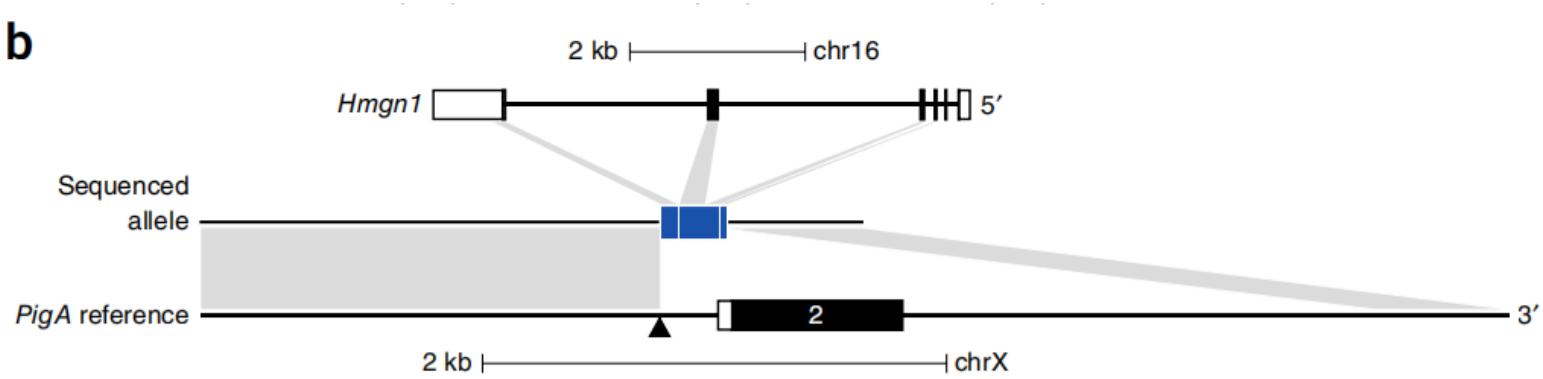
CRISPR-Cas9 leads to complex SVs



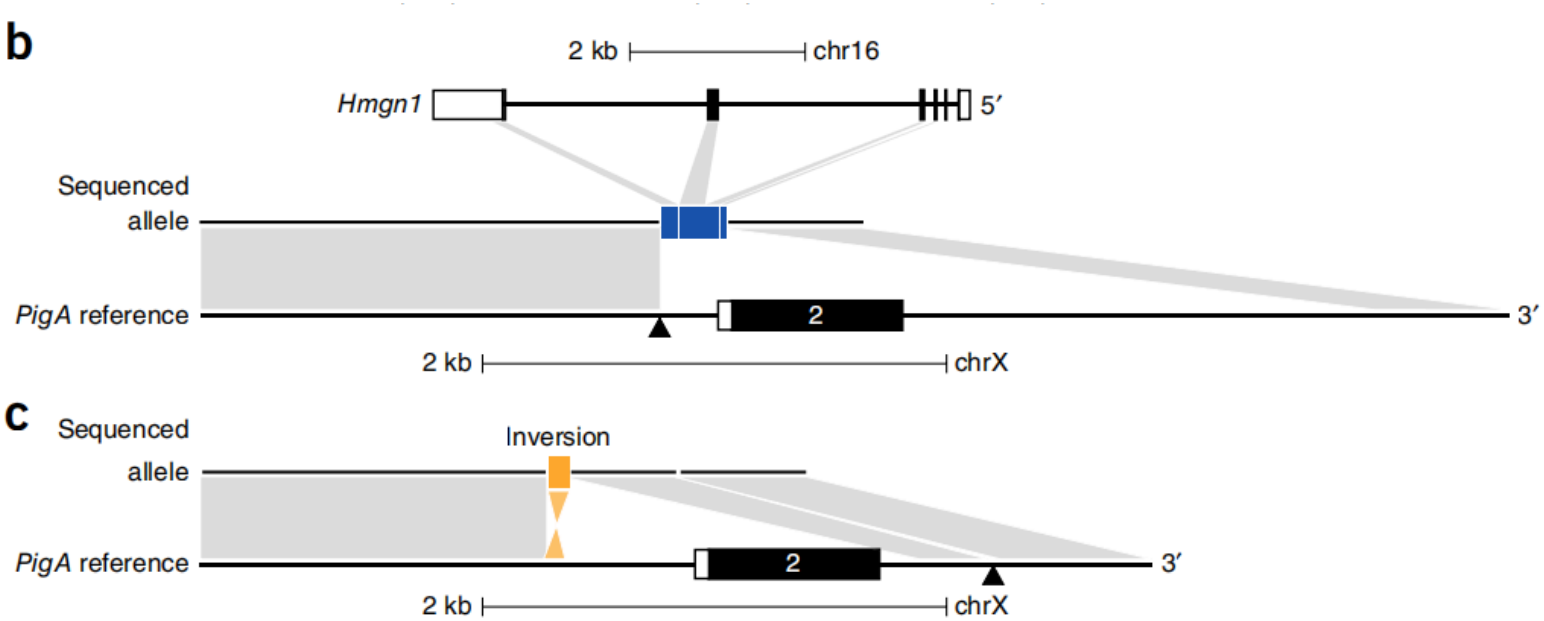
CRISPR-Cas9 leads to complex SVs



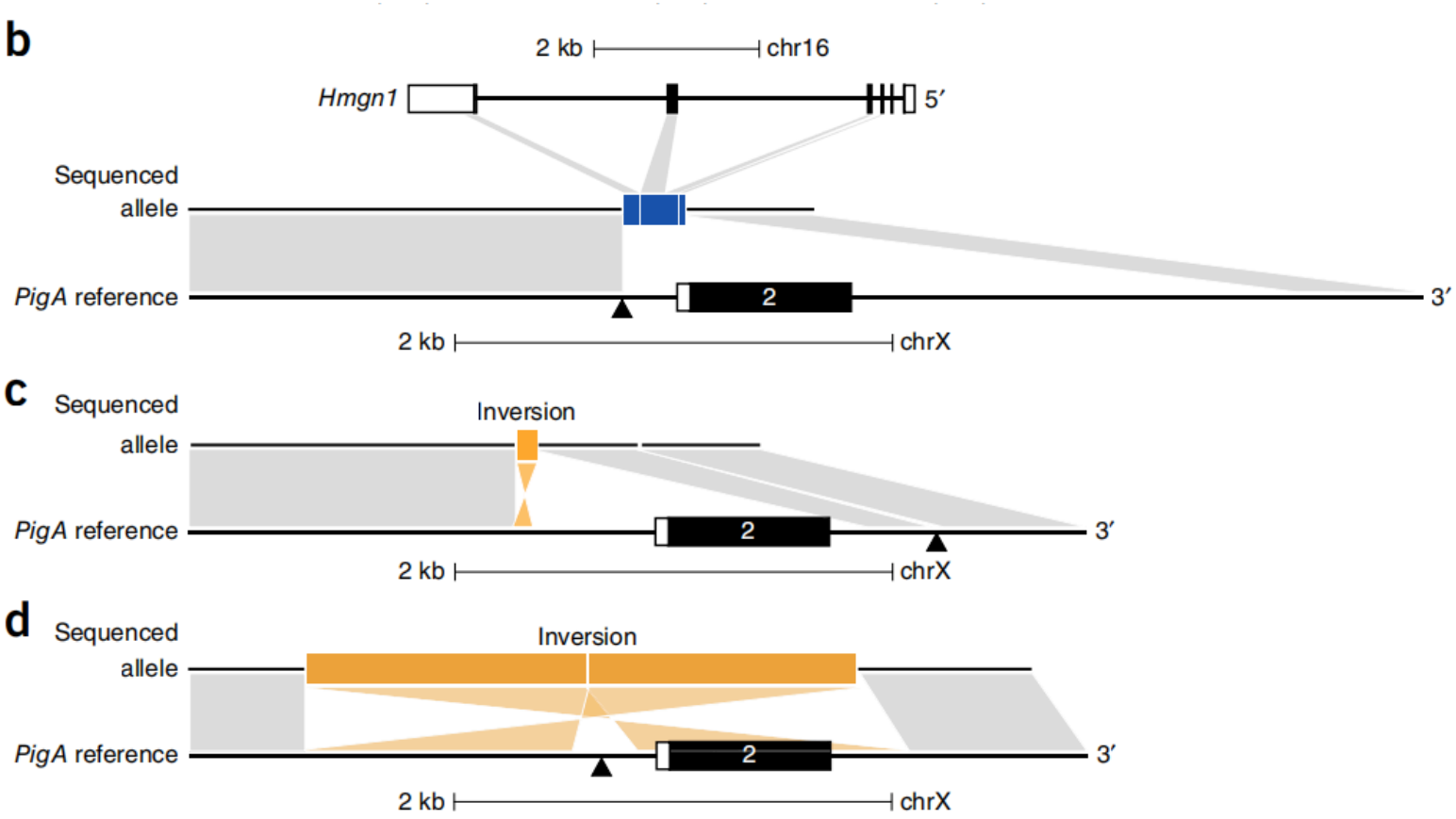
CRISPR-Cas9 induced SVs



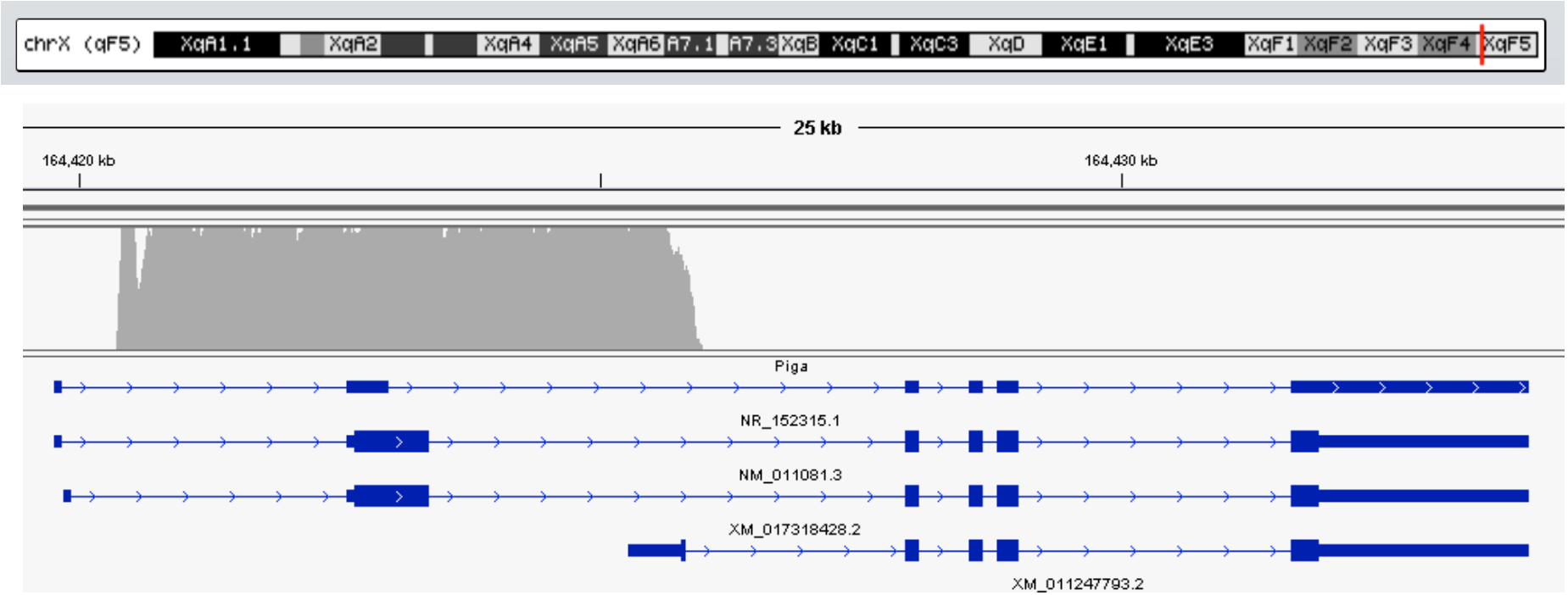
CRISPR-Cas9 induced SVs



CRISPR-Cas9 induced SVs



Targeted sequencing of PigA gene



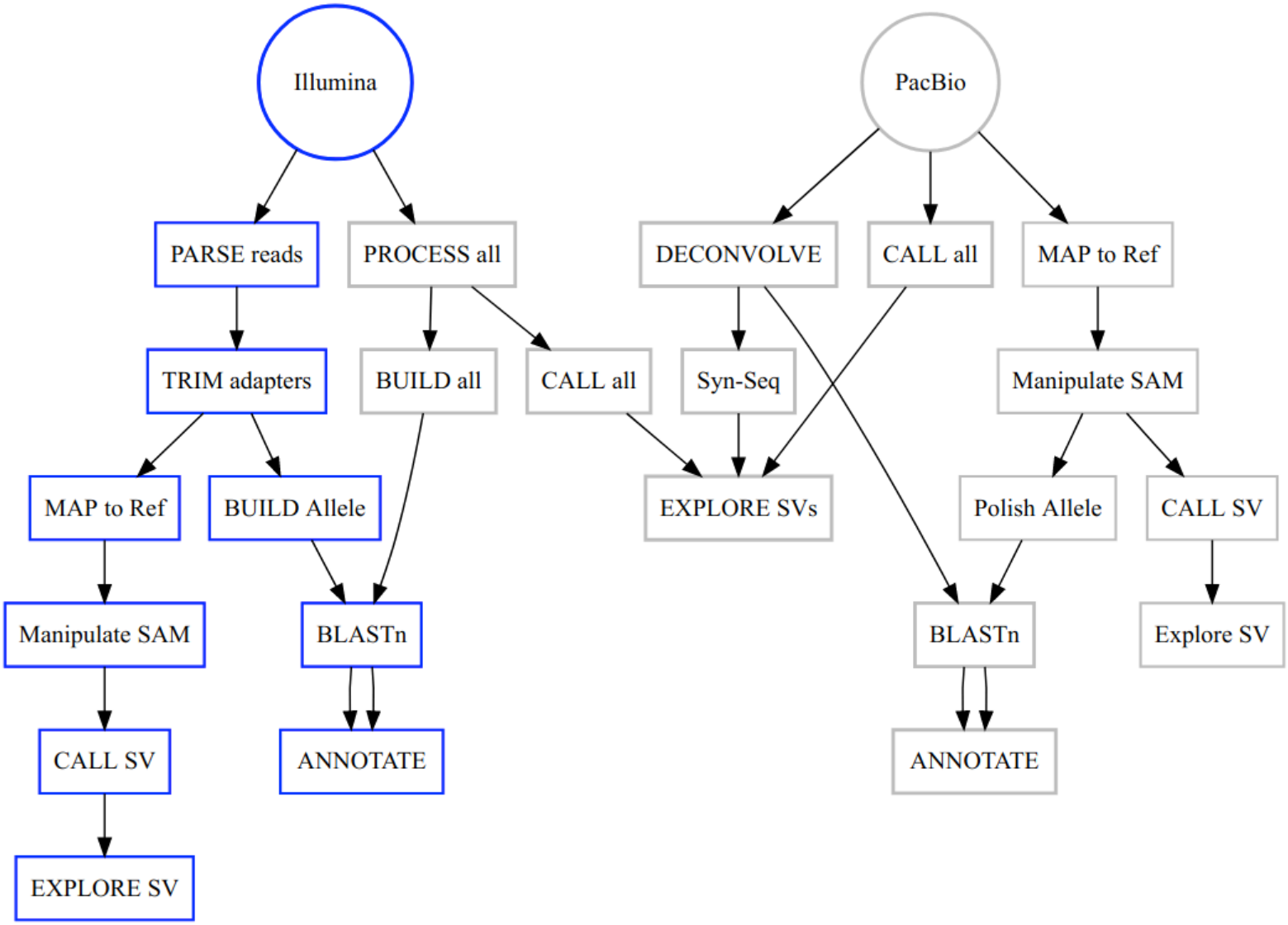
SV_Dataset

- Instructions
  - Ground_Truth.txt
  - ILL_sample.txt
  - PAC_sample.txt
  - PIPE1_tutorial.html
  - PIPE1_workflow.html
  - PIPE2_tutorial.html
  - PIPE2_workflow.html
  - PIPE3_workflow.html
- PIPE1
  - for_the_braves
  - adapters.fa
  - PIPE1_fastq.tar.gz
- PIPE2
  - for_the_braves
  - PIPE2_fastq.tar.gz
- Ref_genome
  - GRCm38.p6_ChrX.fa
  - ref_GRCm38.p6_top_level.gff3.gz

PIPE1: Illumina PE-reads

sample_ID	target	expr	gRNA	library	reads	COV	barcode
PIGA00ILL001	NA	POS	sham	sc	1866	52.1	CTGCGTGCTCTACGAC
PIGA56ILL070	EXON	NEG	g56	sc	1930	54.5	CATAGCGACTATCGTG
PIGA10ILL066	INTRON	POS	g10	sc	2070	58.4	GCTCGACTGTGAGAGA
PIGA15ILL071	INTRON	POS	g15	sc	1740	48.7	ACTCTCGCTCTGTAGA
PIGA15ILL010	INTRON	NEG	g15	sc	1638	65.2	CAGTGAGAGCGCGATA
PIGA15ILL003	INTRON	NEG	g15	sc	1826	70.2	GCAGACTCTCACACGC
PIGA15ILL012	INTRON	NEG	g15	sc	1832	80.7	GTGTGAGATATATATC
PIGA15ILL017	INTRON	NEG	g15	sc	2050	76.6	GACAGCATCTGCGCTC
PIGA15ILL031	INTRON	NEG	g15	sc	1822	63.4	TACTAGAGTAGCACTC
PIGA15ILL050	INTRON	NEG	g15	sc	960	52.8	TGTGTATCAGTACATG
PIGA15ILL059	INTRON	NEG	g15	sc	1500	53.7	ACACGCATGACACACT
PIGA15ILL062	INTRON	NEG	g15	sc	1424	47.0	GATCTCTACTATATGC
PIGA15ILL067	INTRON	NEG	g15	sc	2356	70.3	ACAGTCTATACTGCTG
PIGA15ILL069	INTRON	NEG	g15	sc	2334	72.4	ATGATGTGCTACATCT
PIGA15ILLPOL	INTRON	NEG	g15	poly	16416	442.6	TGCTCGCAGTATCACA

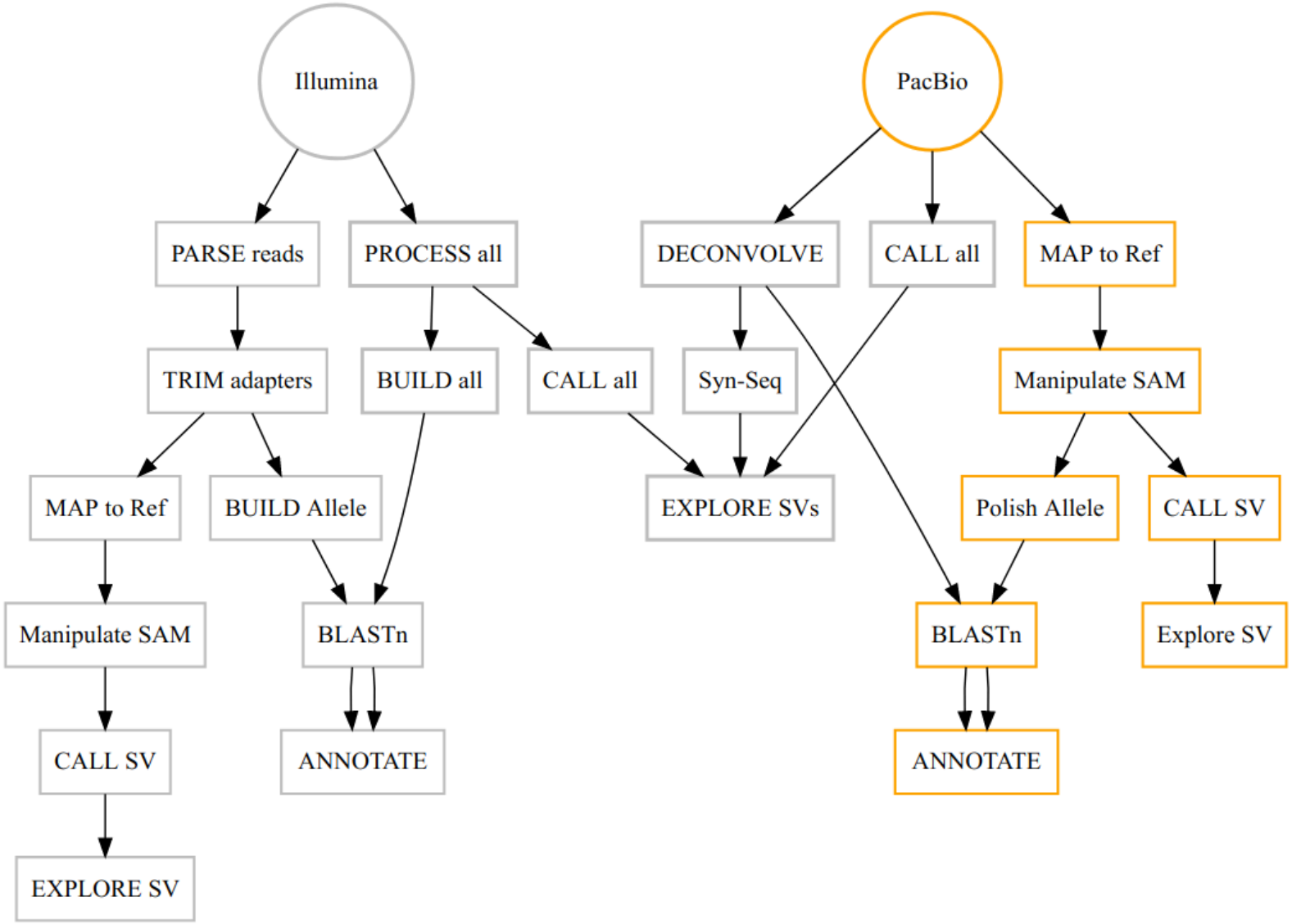
PIPE1: workflow



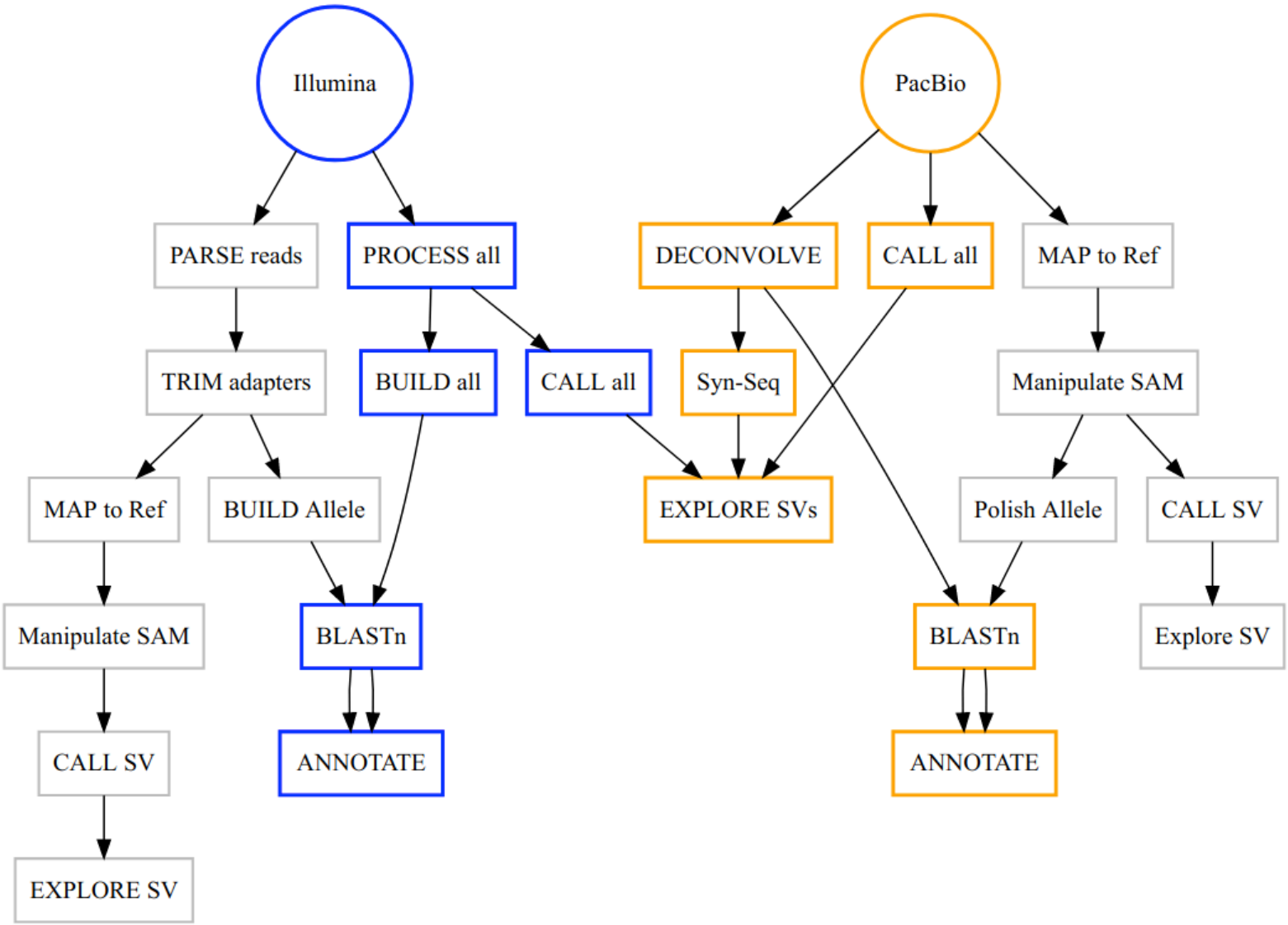
PIPE2: PacBio FLNC reads

sample_ID	target	expr	gRNA	library	reads
PIGA00PAC001	NA	POS	sham	sc	21
PIGA56PAC070	EXON	NEG	g56	sc	21
PIGA10PAC066	INTRON	POS	g10	sc	21
PIGA15PAC071	INTRON	POS	g15	sc	21
PIGA15PAC010	INTRON	NEG	g15	sc	21
PIGA15PAC003	INTRON	NEG	g15	sc	21
PIGA15PAC012	INTRON	NEG	g15	sc	21
PIGA15PAC017	INTRON	NEG	g15	sc	21
PIGA15PAC031	INTRON	NEG	g15	sc	21
PIGA15PAC050	INTRON	NEG	g15	sc	21
PIGA15PAC059	INTRON	NEG	g15	sc	21
PIGA15PAC062	INTRON	NEG	g15	sc	21
PIGA15PAC067	INTRON	NEG	g15	sc	21
PIGA15PAC069	INTRON	NEG	g15	sc	21
PIGA15PACPOL	INTRON	NEG	g15	poly	210

PIPE2: workflow



PIPE3: workflow



Variants Deconvolution